Homework 7

Due: Thursday, March 13, 2025 at 12:00pm **(Noon)**

Problem 1: VC Dimension

(20 points)
For any hypothesis class $\mathcal{H}$ on domain $\mathcal{X}$, to show that the VC Dimension of $\mathcal{H}$ is d, you should prove each of the following:

- There exists a set $C \subset \mathcal{X}$ of size d such that $\mathcal{H}$ shatters C. (Recall that a set is a collection of unique elements.)
- There does not exist a set $C' \subset \mathcal{X}$ of size $d+1$ such that $\mathcal{H}$ shatters C'.

1. Compute and prove the VC dimension for the following hypothesis classes:

   a. The class of signed intervals in $\mathbb{R}$, $\mathcal{H} = \{h_{a,b,s} : a \leq b, s \in \{-1, 1\}\}$ where:

   $$h_{a,b,s} = \begin{cases} s & \text{if } x \in [a, b] \\ -s & \text{if } x \notin [a, b] \end{cases}$$

   b. The class of origin-centered spheres in $\mathbb{R}^d$, $\mathcal{H} = \{h_{a,s} : s \in \{-1, 1\}, a \in \mathbb{R}\}$ where:

   $$h_{a,s} = \begin{cases} s & \text{if } x \text{ is within or on the origin centered sphere of radius } a \\ -s & \text{if } x \text{ is outside the origin centered sphere of radius } a \end{cases}$$

2. Consider two hypothesis classes $\mathcal{H}_1$, $\mathcal{H}_2$ such that $\mathcal{H}_1 \subset \mathcal{H}_2$. Prove that the VC Dimension of $\mathcal{H}_2$ is at least as large as the VC Dimension of $\mathcal{H}_1$.

Problem 2: VC Dimension and PAC Learning

(20 points)

Decision trees can split on data with binary features ($\mathcal{X} = \{0, 1\}^d$) or continuous features ($\mathcal{X} = \mathbb{R}^d$). Assume that the nodes of a continuous decision tree have splitting rules that threshold the value of a single feature. *Note that for continuous decision trees, multiple splits can be made on the same feature. For binary decision trees, only a single split can be made on a feature.*

Consider the following hypothesis classes:

$$\mathcal{H}_1 = \{h : h \text{ is a decision tree for data with only binary features}\}$$
$$\mathcal{H}_2 = \{h : h \text{ is a decision tree for data with only continuous features}\}$$

1. Compute and prove the VC dimension of $\mathcal{H}_1$.
2. Show that the VC dimension is infinite for $\mathcal{H}_2$.
3. Is $\mathcal{H}_1$ PAC learnable? How about $\mathcal{H}_2$? Explain.

Problem 3: Uniform Convergence

(10 points)

In class and in Corollary 4.6 in the textbook, we proved that finite hypothesis classes enjoy uniform convergence, and therefore are agnostically PAC learnable. In those proofs, we assumed that the loss function has a range of $[0, 1]$. Prove that if the range of the loss function is instead $[a, b]$, then the sample complexity satisfies:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon/2, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)(b-a)^2}{\epsilon^2} \right\rceil .$$

You may not use the result from in class or Corollary 4.6 as an argument for your proof, but you may (and it is recommended!) use the same proof steps in guiding your approach.

Problem 4: Socially Responsible Computing: Data Privacy

(5 points)

In 2012, Minerva High School, a public school in Pittsburgh, PA with nearly 3,000 students, hit a record student dropout rate of nine percent.[1] The school principal and board decided to put the extensive data the school had already collected about its students' behavior to use. These datasets included demographic information, academic performance, disciplinary and attendance records, and teacher statistics (i.e. percent of students failing per class, years of teaching). The school also tracked students' internet use and monitored their movements throughout the campus.

The board members suggested that developments in machine learning could be applied to this information to understand what causes students to drop out so that new incentive structures for teachers and students could be created. They contracted a local data science company, Hephaestats, to provide them with their existing databases and gave them access to new data as it was collected. Given the urgency of the situation, the principal proceeded quickly, without time to notify students and parents of this agreement, nor giving them the opportunity to opt out. They justified that this decision was supported by the school board and fell within the general mandate to promote positive educational outcomes for all.

1. In this case, the decision to adopt AI technologies came from above—a suggestion from the school board, implemented by the principal. Who are the other relevant stakeholders, and how could they have been involved? Should they have been involved in the decision to use Hephaestats?

2. Review the introduction of this section on Data Privacy in the Blueprint for an AI Bill of Rights. According to the blueprint, did the school violate the privacy of its students by sharing their data with Hephaestats? If you were the principal, what would you have done?

Grading Breakdown

The grading breakdown for the assignment is as follows:

Problem 1	40%
Problem 2	40%
Problem 3	15%
Problem 4	5%
Total	100%

[1] Case study by Princeton Dialogues on AI and Ethics licensed under CC Attribution 4.0 International

Handing In

You will turn in your final handin via Gradescope, as detailed in the email sent to the course. If you have questions on how to set up or use Gradescope, ask on Edstem! For this assignment, you should have written answers for Questions 1, 2, 3, and 4.

Anonymous Grading

You need to be graded anonymously, so do not write your name anywhere on your handin.

Obligatory Note on Academic Integrity

Plagiarism — don't do it.

As outlined in the Brown Academic Code, attempting to pass off another's work as your own can result in failing the assignment, failing this course, or even dismissal or expulsion from Brown. More than that, you will be missing out on the goal of your education, which is the cultivation of your own mind, thoughts, and abilities. Please review this course's collaboration policy and if you have any questions, please contact a member of the course staff.